[to appear in: Trousdale, Graeme and Thomas Hoffmann (eds.). *The Oxford handbook of construction grammar*. Oxford University Press.]

Words and Idioms

Stefanie Wulff

1. Idioms are constructions

1.1 Idioms in the Chomskyan paradigm

Croft and Cruse (2004:225) describe the role of idioms in construction grammar as follows: "It is not an exaggeration to say that construction grammar grew out of a concern to find a place for idiomatic expressions in the speaker's knowledge of a grammar of their language". The reason for this focus on idioms in early constructionist research is that idioms are incompatible with a theory of language that assumes a strict separation of grammar and the lexicon, such as early versions of transformational-generative grammar (Chomsky 1965), the predominant linguistic framework in the 20th century. Chafe (1968) accordingly referred to idioms as an "anomaly in the Chomskyan paradigm".

One the one hand, idioms are not just fixed word combinations because they can be modified both lexically and syntactically. The idiom *walk a tightrope* ('to act very carefully so that you avoid either of two equally bad but contrasting situations') is an example in question. The following attestation from the *Corpus of Contemporary American English*[1] illustrates that an adjective like *legal* can be inserted before the noun and that the syntactic arrangement of verb and object is not fixed either.

(1) As wellness programs and employee surcharges multiply, employers are acutely aware of **the legal tightrope they must walk**.

At the same time, many idioms defy the standard rules of grammar, so they are not assembled in the same fashion as regular phrases. (2) is an example of an idiom so formally fixed as not to license tense variation; (3) is an idiom that cannot undergo syntactic transformations like passivization; conversely, idioms like the one in (4) are restricted to what would be considered a transformationally derived surface structure in transformational-generative

[1] <http://www.americancorpus.org/>.

grammars (example (2) taken from Croft and Cruse 2004:233; example (4) taken from Nunberg et al. 1994:516; glosses added).

(2) a. It takes one to know one. ('Only a person with identical character traits is able to recognize those traits in someone else.')
b. *It took one to know one.

(3) a. He shot the breeze. ('He engaged I idle conversation.')
b. *The breeze was shot (by him).

(4) a. The dice is cast. ('An irrevocable decision has been made.')
b. *X cast the dice.

1.2 Idioms in the constructicon

Fillmore, Kay and O'Connor (1988) addressed this issue by suggesting that idioms should be seen as units of syntactic representation that are associated with unique functional (semantic/pragmatic) properties. To make their case, they laid out the semantic and syntactic irregularities of the *let alone*-construction as in (5) and (6). Syntactically, *let alone* basically functions like a coordinating conjunction, but it does not license the same syntactic arrangements (examples from Fillmore et al. 1988:515-516):

(5) a. Shrimp and squid Moishe won't eat.
b. *Shrimp let alone squid Moishe won't eat.

(6) a. *Shrimp Moishe won't eat and squid.
b. Shrimp Moishe won't eat, let alone, squid.

Similarly, *let alone* shares some contexts with comparative *than*, but does not license VP ellipsis like *than* does (examples taken from Fillmore 1988 et al.:517, 516):

(7) a. John hardly speaks Russian let alone Bulgarian.
b. John speaks better Russian than Bulgarian.

(8) a. Max will eat shrimp more willingly than Minnie will.
b. Max won't eat shrimp but Minnie will.

c. *Max won't eat shrimp let alone Minnie will.

Moreover, *let alone* can be understood as a negative polarity item, which makes it similar to *any*, but *let alone* is allowed in contexts in which *any* and other negative polarity items cannot occur (examples taken from Fillmore et al. 1988:518-519):

(9) a. He didn't reach Denver, let alone Chicago.
b. He didn't reach any major city.
(10) a. You've got enough material there for a whole semester, let alone a week.
b. *You've got enough material for any semester.

With regard to its semantics, *let alone* requires a complex chain of interpretative processes on the part of the hearer that Croft and Cruse (2004) summarize as follows:

> First the interpreter must recognize or construct a semantic proposition in the fragmentary second conjunct that is parallel to the proposition in the first full conjunct. … More specifically … [t]he interpreter must construct a scalar model, which ranks propositions on a scale – for example, the distastefulness of eating seafood … the initial, full conjunct denotes the proposition that is stronger or more informative on the scale … This whole semantic apparatus is required for the interpretation of the *let alone* construction, and is not necessary (as a whole) for other constructions. (Croft and Cruse 2004:239)

Going beyond the individual example of *let alone*, Fillmore et al. (1988:506-10) classified idioms into different types depending on the extent to which they deviate from regular syntactic expressions in terms of their lexical, semantic, and syntactic irregularity. Table 1 provides an overview.

Table 1. Types of idioms compared to regular syntactic expressions.

	Lexis	**Syntax**	**Semantics**
unfamiliar pieces unfamiliarly arranged	irregular	irregular	irregular
familiar pieces unfamiliarly arranged	regular	irregular	irregular
familiar pieces familiarly arranged	regular	regular	irregular
regular syntactic expressions	regular	regular	regular

When parts of an expression are not found outside the idiom they occur in and the idiom has a syntactically irregular configuration, Fillmore et al. classify the expression as "unfamiliar pieces unfamiliarly arranged"; examples here are *kith and kin* ('family and friends') and *with might and main* ('with a lot of strength'). As Croft and Cruse (2004:235) note; "[u]nfamiliar words are by definition unfamiliarly arranged: if the words do not exist outside the idiom, then they cannot be assigned to a syntactic category in terms of a regular syntactic rule". At the same time, it is important to note that while these expressions may be quite irregular across all three relevant domains, they are not necessarily entirely non-compositional: parts of the expression can still be mapped onto parts of its meaning. We must conclude that even within this group of highly irregular expressions, there is considerable variation as far as the degree of non-compositionality is concerned, with the majority of expressions being partially compositional. That is, most expressions even in this category are not *decoding idioms* (in Fillmore et al.'s terminology) that need to be learned as wholes, but *encoding idioms*.

The category of "familiar pieces unfamiliarly arranged" is home to expressions like *all of a sudden* (*sudden* does not function as a noun outside of this idiom) and *in point of fact* (in other contexts, *fact* requires a determiner). They differ from the first group only in that all their component words are familiar, that is, also occur outside of the expression. They may also vary in terms of their degree of schematization (Fillmore et al. refer to the lexically fully specified idioms as *substantive* and partially lexically specified idioms as *formal idioms*), and to what extent the contributions that the component words make to the expression overlap with their meanings outside of that expression.

The third class, "familiar pieces familiarly arranged", covers (again both substantive and formal) expressions like *pull X's leg* (with X being a lexically unspecified slot that can be filled with a human referent) or *tickle the ivories* ('play the piano'). These differ from the other types of idiomatic expressions only such that their component words "are arranged in a way that reflects the regular grammatical patterns of the language" (Evans and Green 2006:645).

A series of influential studies on schematic constructions that are semantically and syntactically more or less irregular followed Fillmore et al.'s (1988) case study on *let alone*. Moreover, various earlier studies within the larger family of cognitive-functional linguistics were reassessed from a specifically constructionist perspective.[2] Table 2 provides an overview of some of these constructions, example sentences, and references (in chronological order of publication). The selection of studies listed here is by no means representative of the vast (and steadily growing) number of studies published from, or at least compatible with, a constructionist perspective, but illustrates the range of constructions that can be described by means of parameters of complexity, lexical specification, and semantic and syntactic irregularity.

[2] It is also important to point out here that outside of (direct predecessors of) construction grammar, there is vast amount of literature on idiomatic language, some of it considerably predating constructionist work, yet resonating in essence with the fundamental claims made in construction grammar. Examples here include (mainly European) phraseological research (see Cowie and Howarth (1996) provide a select bibliography), discourse-analytical approaches (Pawley and Syder's (1983) "formulas", Hymes' (1962) "conversational routines", or Nattinger and DeCarrico's (1992) "formulaic sequences" are but three examples of different notions of idiomatic structures), and a plethora of psycho-linguistic work (Cacciari and Tabossi (1995) may be a good starting point here). The reader should furthermore be referred to Wray (2002), who gives an excellent overview of phraseological research and its implications for theories of the mental lexicon.

Table 2. Examples of semantically and/or syntactically irregular constructions.

Construction label	**Example (reference)**
omitted determiners	*I don't think, Mac/*cabby, that this is the best way to go*. (Zwicky 1974)
it-clefts	*It is against pardoning these that many protest*. (Prince 1978)
deictic *there* construction	*There goes the bell now!* (Lakoff 1987)
tautologies	*Boys will be boys*. (Wierzbicka 1987)
syntactic amalgams	*There was a farmer had a dog* (Lambrecht 1988)
have/give/take a V	*have a drink*; *give the rope a pull*; *take a look at* (Wierzbicka 1988)
mad magazine construction	*Him, a doctor?!* (Lambrecht 1990)
N P N construction	*house by house*; *day after day* (Williams 1994)
nominal extraposition	*It's amazing the difference!* (Michaelis and Lambrecht 1996)
time *away* construction	*Twistin' the night away* (Jackendoff 1997)
preposing	*It's very delicate, the lawn.* (Birner and Ward 1998)
What's X *doing* Y?	*What's that fly doing in my soup?* (Kay and Fillmore 1999)

Following Fillmore et al.'s reconceptualization of idioms as symbolic units and the ensuing wealth of studies on constructions, "construction grammarians came to argue that, in fact, grammatical organization is entirely vertical" (Croft and Cruse 2004:247-8). For a schematic representation of this "vertically organized" extended mental lexicon, or so- called "constructicon", please see Goldberg (this volume). As Goldberg explains, the difference between words and grammatical frames is one of degree rather than quality, and one can describe it along two parameters, complexity and schematization. Firstly, constructions differ in terms of their complexity: morphemes and words are simple constructions, whereas idioms and grammatical frames are increasingly complex. Secondly, constructions differ in their degree of schematization or lexical specification: words are fully lexically specified, whereas grammatical frames are maximally unspecified with regard to the lexical material that can be inserted. Idioms

occupy the space in between these two extremes, with some like *shoot the breeze* being fully lexically filled and others like *pull X's leg* being only partially specified.[3]

1.3 Idioms in usage-based construction grammar

As we have seen in the section above, the integration of semantically and syntactically irregular expressions, first and foremost idioms, has been a major impetus for the development of construction grammar. It must be emphasized here, however, that formal and/or functional unpredictability are sufficient, but not necessary conditions for construction status: even highly regular expressions that are used sufficiently often to become entrenched in the speaker's mental lexicon qualify as constructions (Goldberg 2006:64; see also Bybee, this volume). This qualification of earlier definitions of constructions (Goldberg 1995:4) as necessarily irregular is being promoted most strongly in usage-based construction grammar (see Goldberg, this volume). In usage-based construction grammar, the acquisition, representation, and processing of language are shaped by usage. Methodologically, a usage-based perspective entails (among other things) a decided focus on (ideally representative samples of) authentic language data. As such, usage-based construction grammar reflects the general empirical turn in linguistics (see Gries, this volume).

The inclusion of expressions in the constructicon based on frequency alone does not contradict the hypothesis that it is primarily vertically arranged. However, it strongly emphasizes the idea that different factors like lexical specification, semantic irregularity, syntactic irregularity, and cognitive entrenchment prevail at all levels of the constructicon in *different shades of prominence and relative importance.* It appears that at the level of idioms (in particular not fully lexically specified ones), we not only find all of these factors instantiated, which makes schematic idioms particularly interesting; what is more, any given (schematic) idiom can be characterized individually along every single factor, resulting in a "multi-dimensional continuum" of differently formally and semantically irregular and cognitively entrenched expressions that ultimately blurs the boundaries of idiom types as described in Fillmore et al.

[3] The concept of "constructional idioms" has also been applied to languages other than English. For a treatment of particle verbs in German, for instance, see Booij (2010), which provides plenty of examples.

(1988) and various other, non-constructionist idiom typologies (e.g. Fernando 1996; Cacciari and Glucksberg 1991). Moreover, such an understanding renders the term "irregularity" rather misleading; instead, I will henceforth refer to the semantic and formal "behavior" of constructions.

From a usage-based construction grammar perspective, it begs the question how to describe this continuum most adequately. In what follows, I will present measures derived from corpus data for the semantic and syntactic irregularity of differently frequent substantive V NP constructions that stand in accord with basic constructionist premises. I will then discuss which of these factors, and to what extent, underlie native speaker judgments of the perceived idiomaticity of these constructions.

The data in question are 39 V NP constructions (13,141 tokens total) retrieved from the British National Corpus (BNC). 33 of these constructions are listed in the *Collins Cobuild Idiom Dictionary*, and six constructions were added based on frequency alone (see Wulff 2008: 25-27 for details). (11) lists all 39 V NP constructions with their frequencies (in all their variant forms) in parentheses.

(11) *bear* DET[4] *fruit* (90), *beg* DET *question* (163), *break* DET *ground* (133), *break* DET *heart* (183), *call* DET *police* (325), *carry* DET *weight* (157), *catch* DET *eye* (491), *change* DET *hand* (212), *close* DET *door* (827), *cross* DET *finger* (150), *cross* DET *mind* (140), *deliver* DET *good* (145), *do* DET *trick* (155), *draw* DET *line* (310), *fight* DET *battle* (192), *fill/fit* DET *bill* (116), *follow* DET *suit* (135), *foot* DET *bill* (109), *get* DET *act together*[5] (142), *grit* DET *tooth* (164), *have* DET *clue* (232), *have* DET *laugh* (98), *hold* DET *breath* (292), *leave* DET *mark* (145), *make* DET *headway* (136), *make* DET *mark* (213), *make* DET *point* (1,005), *make/pull* DET *face* (371), *meet* DET *eye* (365), *pave* DET *way* (269), *play* DET *game* (290), *scratch* DET *head* (100), *see* DET *point* (278), *take* DET *course* (294), *take* DET *piss* (121), *take* DET *plunge* (115), *take* DET *root* (113), *tell* DET *story* (1,942), *write* DET *letter* (1,370)

[4] DET stands for any kind of determiner, including a zero determiner.

[5] While this is not a V NP construction, it was included in several pre-tests and is reported alongside the V NP constructions.

2. The contribution of words to idioms: measuring semantic behavior

As discussed above, constructionist studies of idioms agree that "the meanings of most idioms have identifiable parts, which are associated with the constituents of the idioms" (Nunberg et al. 1994:531; see also Lakoff (1987:448ff.) for a detailed account of how metaphor theory can be employed to account for the linkages between literal and idiomatic meaning). At the opposite end of the compositionality spectrum, we likewise find that "[s]trict compositionality is rarely, if ever, encountered. Most expressions (I am tempted to say: *all* expressions), when interpreted in the context in which they are uttered, are non-compositional to some degree" (Taylor 2002:550). In other words, a cognitive-linguistic perspective entails that compositionality is not a binary, but a scalar concept.

This view is supported by findings from various psycholinguistic studies which suggest that the literal meanings that are activated during processing facilitate idiomatic construction comprehension to the extent that they overlap with the idiomatic meaning. For instance, Gibbs and colleagues (Gibbs and Nayak 1989; Gibbs et al. 1989) demonstrated that subjects can distinguish between at least three classes of idiomatic constructions in terms of their compositionality, and that sentences containing decomposable constructions are read faster than those containing non-decomposable constructions (see also Glucksberg 1993; Peterson and Burgess 1993; Titone and Connine 1994; McGlone et al. 1994).

A constructionist perspective furthermore entails a number of working assumptions that an adequate compositionality measure should be able to incorporate. Firstly, any complex construction is assumed to comprise a number of smaller constructions, all of which make a semantic contribution to that complex construction (in other words, constructions further up in the constructicon as shown in Figure 1 feed into the semantics of constructions further down of which they are part; see Goldberg 2006:10). In the case of V NP constructions, both the verb and the noun phrase of a V NP construction are expected to make a contribution.

Secondly, as mentioned above, constructions are assumed to be differently entrenched in the constructicon depending on (among other things) their frequency of use. Accordingly, a theoretically informed measure should license the possibility that component words make variably large contributions (since there is no reason to assume that verbs and noun phrases necessarily make equally large contributions to V NP constructions). Moreover, the measure should be item-specific in the sense that the contribution of any component word can be

differently large depending on the construction in which it occurs. For instance, it is desirable to have a measure that licenses the possibility that the contribution made by *point* in *see a point* is higher or lower than in *make a point*.

A number of compositionality measures have been proposed that start out from the assumption that semantic regularity is a function of the semantic similarity of the constituent words and the phrasal expression. Some measure compositionality via the ability to replace component words without losing the idiomatic interpretation of the construction. For instance, McCarthy, Keller and Carroll (2003) automatically extracted verb-particle constructions from the BNC. They then retrieved the semantic neighbors of these constructions from an automatically acquired thesaurus. The relative compositionality of a given verb-particle construction was defined as the overlap of the semantic neighbors and the component words of the construction (see also Lin 1999 for another substitution-based measure).

Other approaches, including the one presented here, measure compositionality via the semantic similarity of the contexts of the constructions compared with those of its component words. More specifically, the working hypothesis is that the semantic similarity of two words or constructions is reflected in the extent to which they share collocates. Collocates of words are "the company they keep", that is, words that occur in a (usually user-defined) context window left or right of the word more often than would be predicted on the basis of the word's general frequency. The more semantically similar two words or constructions are, the more similar their contexts will be. To give an example of a statistically sophisticated measure along those lines, Schone and Jurafksy (2001) extracted multiword-expressions (MWEs) from corpora using Latent Semantic Analysis. The compositionality of these MWEs was measured as the cosine between the vector representation (containing collocation frequencies) of the MWE and a weighted vector sum of its component words, the assumption being that small cosines indicate compositionality. (For other measures based on semantic similarity, see e.g. Bannard, Baldwin and Lascarides 2003; Bannard 2005).

The measure to be presented here in a little more detail is an extension of Berry-Rogghe (1974) called weighted *R*. In a first step, the sets of significant collocates for each component

word W and the construction C they are part of are determined.[6] Two index values are then combined to arrive at a word's overall contribution to the construction: *R* (which is the number of collocates shared between the word and the construction divided by the total number of collocates of the construction); and the word's "share" (which is the number of collocates shared between the word and the construction divided by the total number of collocates of the word). Both indices can range between 0 (if there is no overlap at all) and 1 (if the collocate sets match perfectly).

(12) $$\text{weighted } R_{\text{W}} = R \times \text{share} = \frac{\text{n collocates W} \cap \text{C}}{\text{n collocates C}} \times \frac{\text{n collocates W} \cap \text{C}}{\text{n collocates W}}$$

By combining the two index values, weighted *R* assesses the overall compositionality of a construction from two complementary perspectives: *R* reflects how much of the semantics of the construction is accounted for by the component word; conversely, the share reflects how much of *itself* each component words brings into the constructional meaning.

In order to illustrate the motivation for this approach, consider the V NP construction *make* DET *mark*. Obviously, *make* is a high-frequency verb, and since the number of significant collocates a word will attract is naturally correlated with its overall frequency, *make* has many significant collocates. The noun *mark*, on the contrary, is much less frequent, and consequently, it attracts fewer significant collocates. In sum, the collocate sets of *make* and *mark* differ in size considerably. From this, we can deduce that *make* stands a much higher chance to contribute to *any* construction's semantics than *mark*; what is more, since lexically fully specified complex constructions cannot be more frequent than their component words and accordingly always have comparatively smaller collocate sets, the resulting overlap between a highly frequent component word's collocates and the construction it is part of will be quite high *by default*. In the case of *make* DET *mark*, *R* actually amounts to 1.0: *make* DET *mark* attracts 33 significant collocates, all of which it shares with *make*'s collocate set. In other words, the semantic contribution of *make* to *make* DET *mark* may be considered extremely high when looking only at how much of

[6] Significant collocates were calculated using a Fisher Yates exact (FYE) test (see Stefanowitsch, this volume). Collocates had to yield an association strength of FYE≥100 to enter into a collocate set.

the construction's semantics is accounted for. When we look into the opposite direction, however, we see that *make* contributes only a fraction of its meaning potential: the 33 collocates it shares with *make* DET *mark* constitute only a small share of its total collocate set of 4,234 collocates. This calls for a re-evaluation of the semantic similarity between *make* and *make* DET *mark*. By multiplying the *R*-value with the share value, we achieve exactly that.

For *mark*, a very different picture emerges: the overlap between *mark*'s collocate set and that of *make* DET *mark* is 31, which again indicates a high semantic contribution, and since *mark* attracts 298 collocates overall, the share of 31 out of 298 is relatively high. That is, *mark* is semantically much more similar to *make* DET *mark* than *make* is in the sense that it is much more semantically tied to this construction, while *make* occurs in so many different contexts so much more often that one cannot speak of a particularly tight semantic association between *make* and *make* DET *mark*.

The overall compositionality value of a construction C is defined as the sum of the weighted contributions of all its component words W (in the case of V NP constructions, the verb and the noun phrase). Figure 2 provides an overview of the results for the V NP constructions (the exact values are can be found in Wulff 2009:137).

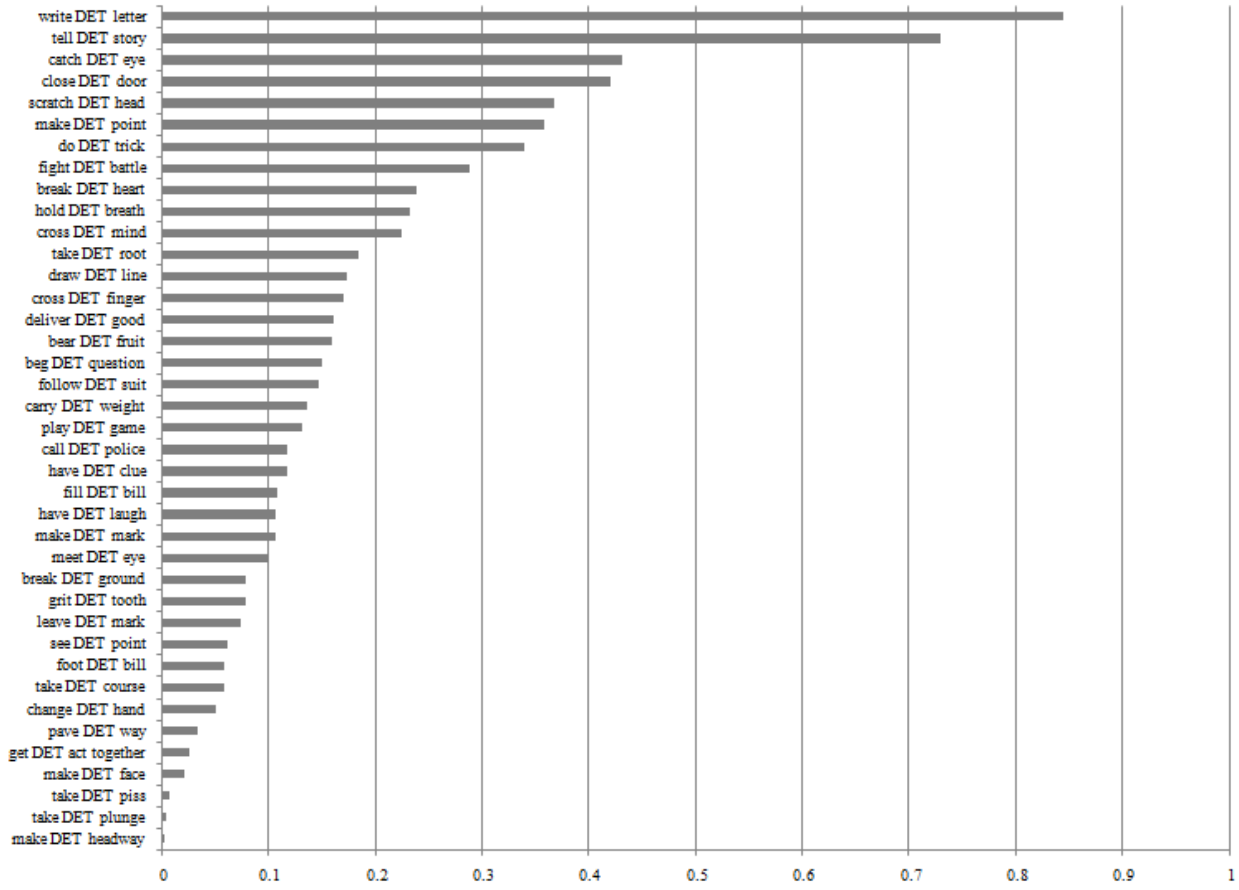

Figure 2. V NP constructions and their weighted *R*-values

As can be seen in Figure 2, the weighted *R*-measure neatly reproduces what we would expect from established idiom typologies: idioms like *make* DET *headway* and *take* DET *plunge* rank lowest in compositionality; metaphorical expressions like *make* DET *mark* and *meet* DET *eye* occupy the middle ranks; quasi-metaphorical constructions, the literal referent of which is itself an instance of the idiomatic meaning, like *cross* DET *finger*, *hold* DET *breath*, and *scratch* DET *head*, tend to rank even higher in compositionality; and most of the constructions that were not picked from the idiom dictionary rank highest, with *write* DET *letter* yielding the highest weighted *R*-value.

Note also that the majority of items is assigned a fairly non-compositional value on the scale from 0 to 1, which ties in nicely with the fact that most of these were actually obtained from an idiom dictionary. Items such as *write* DET *letter* and *tell* DET *story*, on the other hand, were selected to test if items that are intuitively assessed as (nearly perfectly) compositional are

actually treated accordingly by the measure – so weighted *R* proves very accurate since these items do not only rank highest, but moreover, their compositionality values are very high in absolute terms (.73 for *tell* DET *story* and .84 for *write* DET *letter*).

Weighted *R* may be used not only for assessing the compositionality of V NP constructions, but is principally applicable to any kind of construction. Beyond that, weighted *R* may prove a useful tool for quantitative approaches to related issues in construction grammar and cognitive semantics. For instance, it could be employed to quantify the degree of semantic bleaching of verbs as a function of incipient grammaticalization processes.

3. Measuring the formal behavior of idioms

There is comparatively little research to date on the formal behavior of idioms, particularly empirical studies (exceptions are Moon 1998; Nicolas 1995). Moreover, most studies have been concerned specifically with the syntactic flexibility of idioms and mostly disregarded other aspects of formal behavior. Maybe the best-known study on syntactic flexibility is Fraser (1970), who proposed an implicational "frozenness hierarchy" for idioms that comprised six levels of syntactic transformations such as nominalization or particle movement. The validity of this frozenness hierarchy was taken into question by various scholars (McCawley [Dong] 1971 provided a variety of (vulgar) counter-examples, Makkai (1972) even more, if only not vulgar).

However, various psycho-linguistic studies have established a connection between assessments of an idiom's compositionality and its formal properties, including positive correlations between an idiom's compositionality and its rated syntactic flexibility (Gibbs and Nayak 1989), lexical flexibility (that is, regarding the question if and to what extent material can be inserted; Gibbs et al. 1989), and rated semantic productivity (that is, the creation of variant meaning through word substitutions; McGlone et al. 1994).

From a usage-based perspective, a measure of formal behavior should be maximally data-driven in the sense that there are no a priori expectations as to the syntactic configurations, lexical insertions, or morphological flexibility of a given construction; instead, these categories emerge bottom-up by looking at a large sample of corpus data. In their totality, these categories constitute a construction's formal behavioral profile (see Barkema 1994 for such an approach to nominal compounds).

Wulff (2008) developed such a formal behavioral profile by coding all 13,141 instantiations of the afore-mentioned 39 V NP constructions according to any morphological, lexical, or syntactic variation instantiated in the data. From this data-driven approach, 18 variables emerged that measure different aspects of formal behavior, including the syntactic configuration realized, various aspects of lexico-syntactic behavior (such as the presence of adverbs or attributive adjectives preceding the noun), and the morphological variations possible in English, including tense, aspect, person, etc.; Wulff (2009: 151-3) provides an overview of these variables. Each construction's formal behavioral profile was then compared with that of a random sample of 1,151 V NP constructions serving as a baseline. More specifically, for every of the 18 aspects of formal behavior, the information obtained for each variable level was converted into an overall index value by determining the given construction's deviation from the V NP-baseline. To illustrate how the different variable levels were weighted, consider the Table 3, which summarizes the behavioral of *foot* DET *bill* regarding the morphological variable TENSE. The observed frequencies (n_{obs}) for each variable level are compared with the frequencies expected from the V NP baseline (n_{exp}); deviations between observed and expected frequencies are squared (SSD), summed, and normalized (NSSD).

Table 3. *Foot* DET *bill*'s formal behavioral profile for MF_TENSE.

TENSE	**n_{obs} (%)**	**n_{exp} (%)**	**n_{obs}-n_{exp} (%)**	**SSD**	**Summed SSD**	**NSSD**
past	9.17	25.97	-16.80	282.24	1700.625	0.234
present	41.28	61.94	-20.66	426.836		
future	8.26	1.57	6.69	44.756		
non-finite	41.28	10.51	30.77	946.793		

Note how by squaring the deviations of the observed and expected frequencies, we again yield a weighting of the contribution of the different variable levels that stands in accord with constructional premises: small deviations contribute only little to the overall value, while big deviations will contribute much more. Space does not permit a presentation of the results for all 39 constructions and all 18 aspects of formal behavior; an overview table is given in Wulff (2009:154-5), and Wulff (2008) discusses all results in detail.

4. Constructions are idioms

4.1 The relationship between idiomatic variation and idiomaticity

Another question arising from a usage-based perspective on idioms concerns the relationship between a construction's semantic and formal behaviour as it manifests itself in corpus data (in sum, the construction's "idiomatic variation") and native speakers' assessment of a construction's "idiomaticity", which is an inherently psychological construct. A usage-based approach predicts that idiomaticity judgments will be based (at least in part) on the speaker's processing of his or her linguistic environment with regard to the construction's behaviour. So which of these aspects are associated with idiomaticity judgments, and how strongly?

In order to address this question, a multiple regression analysis was computed with the weighted *R*-values of each V NP construction representing its semantic behaviour and the NSSD values for the 18 aspects of formal behaviour representing the constructions' formal behavioural profile as the independent variables, and normed idiomaticity judgments of these constructions as the dependent variable.[7] Taking all variables into account, nearly 57% of the variance in the average idiomaticity judgments is accounted for, a highly significant result that testifies to a solid relationship between the variables and the judgments (adjusted R^2=.565, p=.005**).[8] More specifically, the regression analysis provides so-called beta weights[9] for all variables; the closer a beta weight is to 1, the more important (in the sense of covering variance) it is. Variables with beta weights +.22 can be considered relevant because they account for 5% of the variance. Consider Table 4 for an overview.

[7] 39 first-year students of English at the University of Sheffield were asked to assess the overall idiomaticity of the 39 V NP constructions. Each participant was given a different construction as a reference construction and then asked to judge the idiomaticity of the other 38 constructions relative to this reference construction (for details, see Wulff 2008: 28-33).

[8] While the adjusted R^2-value only amounts to .565, it has to be borne in mind that this value is lowered by the overall number of variables entering into the computation: the more variables are required to account for all the variance in the data, the lower the adjusted R^2 will be.

[9] Beta weights quantify the contribution of each individual independent variable to the overall correlation observed.

Table 4. Beta weights for variation parameters as determined by a multiple regression of corpus and judgment data.

Variable	Abbreviation	Absolute beta weight
number of verb	MF_NUMV	.757
mood of verb	MF_MOOD	.695
kind of adverb inserted	LF_KINDADV	.651
number of adverbs inserted	LF_NOADV	.632
compositionality	COMP	.578
syntactic configuration	SF	.573
voice of verb	MF_VOICE	.351
negation of verb	MF_NEG	.275
material inserted	LF_ADDITION	.265
corpus frequency	CORPFREQ	.209
person of verb	MF_PERSON	.197
gerundial verb	MF_GERUND	.16
tense of verb	MF_TENSE	.125
number of NP	MF_NUMNP	.109
inserted attributive NP	LF_ATTRNP	.083
realization determiner	MF_DET	.055
aspect of verb	MF_ASPECT	.046
inserted PP	LF_PP	.043
inserted relative clause	LF_RELCL	.038
inserted attributive adjective	LF_ATTRADJ	.032

As Table 4 shows, the most important variables are the morphological variables encoding the number of the verb (MF_NUMV) and the mood of the verb (MF_MOOD), followed by two lexico-syntactic variables, LF_KINDADV (which assesses what kind of adverb is present, if any) and LF_NOADV (a count of the number of adverbs realized, if any). Next in line are compositionality and the syntactic configuration in which the construction is realized (SF). The morphological flexibility parameters MF_VOICE (encoding the voice of the verb) and MF_NEG (encoding the verb as either negated or not) also yield sufficiently high beta weights to be considered relevant. The last variable with a value higher than +.22 is the lexico-syntactic variable LF_ADDITION (.265), which counts if the given construction was lexically modified in any way. The construction's corpus frequency (CORPFREQ) obtains a beta weight of .209. These results coincide in a striking fashion with those of a Principal Components Analysis (PCA) of the corpus-based results (see Wulff 2008:150-6).

In sum, speakers seem to rely on a variety of parameters when judging the overall idiomaticity of V NP-constructions, with a particular focus on the morphological and lexico-syntactic variability of the verb. Furthermore, the multifactorial analysis suggests that tree-syntactic and semantic features of the phrase play an important, yet secondary role – a result that stands in opposition to the widely held assumption that compositionality is the most decisive parameter contributing to idiomaticity. More generally, the solid correlations between the corpus-based measures and the judgment data support a performance-based approach to language.

4.2 The extended constructicon

How could this probabilistic and complex information be implemented into existing schematic models of the constructicon? As outlined above, the constructicon is mainly specified with regard to its vertical axis, with delexicalization (or schematization) being the primary process creating diversification along this axis. To a certain extent, idiomatization can be conceptualized as being diametrically opposed to delexicalization: the more idiomatic a construction is, that is, the more formally and semantically irregular, the less likely it is that this construction will delexicalize. In other words, on a continuum of idiomatic phrases ranging from collocations to idioms, the more idiomatic the phrase, the less delexicalization potential it has.

Accordingly, the constructicon could be extended by a horizontal axis as shown in Figure 3 which cuts across the range of the vertical axis where fully lexically specified complex constructions are located. More precisely, one can think of the constructicon as bifurcating beyond the level of words, opening a quadrant space in which constructions can be positioned according to their degree of schematization and idiomaticity. The closer a phrasal construction is located on the horizontal axis to the vertical axis, the more semantically and syntactically regular it is (e.g. *write a letter*); the more formally frozen and semantically opaque a construction is (such as *take the plunge*), the further away from the vertical axis that construction is positioned on the horizontal axis.

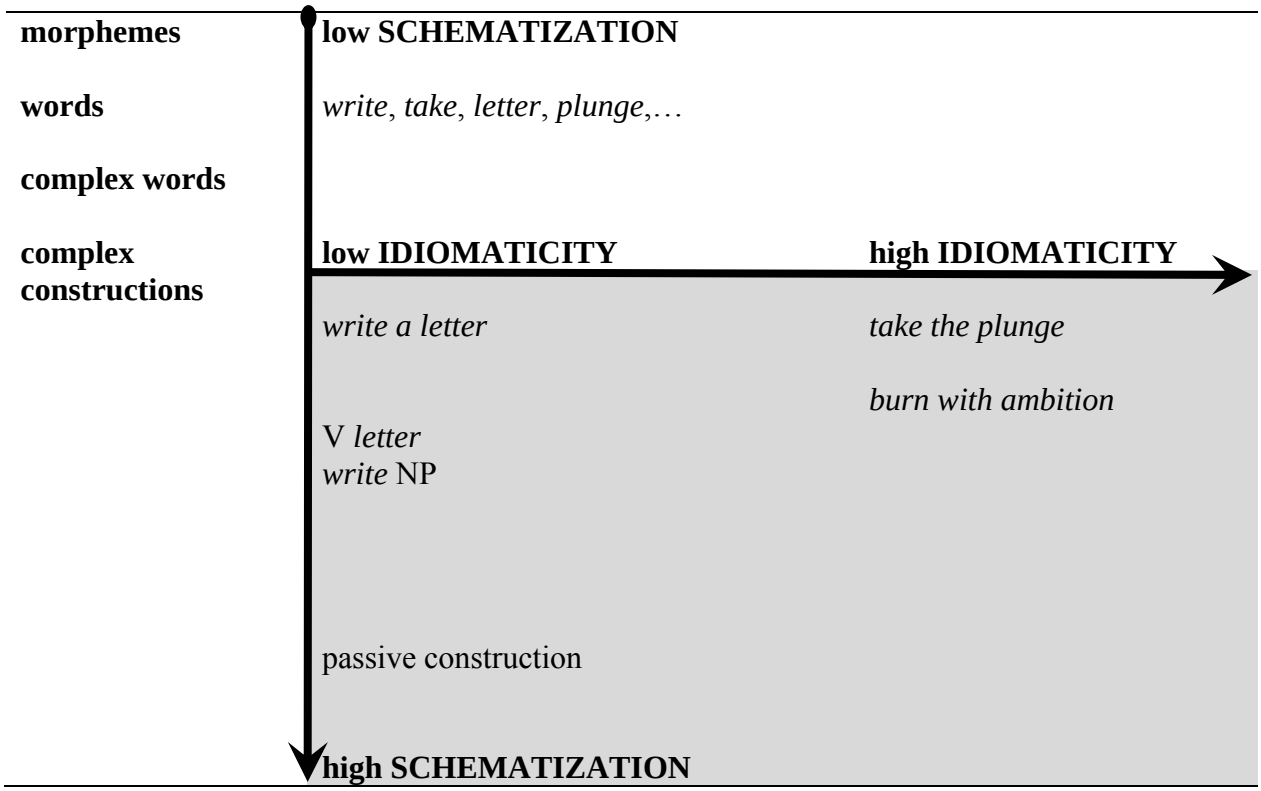

Figure 3. Extended schematic representation of the constructicon

The higher the overall idiomaticity of a construction, the less its representation is connected to its constituting lexical constructions. For instance, *take the plunge* is both semantically highly irregular and formally restricted, so its overall idiomaticity is high; accordingly, *take the plunge* is only weakly connected with the lexical representations of *take* and *plunge*. *Write a letter*, on the other hand, is both formally and semantically regular, so its connection with the lexical representations of *write* and *letter* further up the vertical axis is comparatively stronger. Likewise, it is more strongly connected with other lexical constructions that are associated with *write* and *letter*, such as *type*/*compose* or *email*/*paper*, which in turn makes *write a letter* a likely candidate for subsequent schematization.

5. Conclusions and desiderata

This chapter started out with a summary of early constructionist research that argued in favor of viewing idioms not as anomalies, but constructions that are essentially on a par with all other kinds of constructions. From a usage-based construction grammar perspective, in particular, we

may want to take this statement one step further: all constructions are idiomatic. What may license referring to some constructions as idioms and not others is merely a reflection of the fact that effects of idiomatic variation are *best observable* in partially schematic complex constructions – however, this does not make them *fundamentally* different in nature from other constructions.

Viewing all complex constructions as more or less idiomatic emphasizes the need for more empirical research into the variables that constitute idiomatic variation and their measurement. Speelman et al. (2009) describe the agenda as follows:

> "[A]t least in some cases, constructions are more strongly characterized by the (construction-internal) collocations that instantiate them than by the single items that instantiate them. Consequently, the syntagmatic axis should become a constitutive dimension in a comprehensive Construction Grammar model." (Speelman et al. 2009:87)

An adequate modeling of this syntagmatic axis in Construction Grammar (or what I referred to as the horizontal axis in the constructicon) calls for future research going beyond the studies and measures reviewed here in various regards. As to the semantic behavior of constructions, future studies should systematically explore what kind(s) of measure(s) are suited best to model the association strength between constructions at different levels of schematization. Weighted *R* is just one of many collocation-based association measures that are compatible with constructionist premises, if only highlighting different aspects of the syntagmatic and paradigmatic dimensions of constructional interaction. A first example of such a contrastive analysis is presented by Speelman et al. (2009), who contrast Gries and Stefanowitsch's "Collostructions" (Stefanowitsch and Gries 2003) and "co-varying collexemes" (Gries and Stefanowitsch 2004) for measuring the association strength between different inflectional variants of Dutch attributive adjectives and their head nouns. Another desirable strand of future research could address the relationship between idiomatic variation and idiomaticity in more depth by comparing and validating corpus-derived measures of idiomatic variation in controlled experimental settings.

References

Bannard, C., Baldwin, T., and Lascarides, A. (2003). 'A Statistical Approach to the Semantics of Verb-Particles', in Proceedings of the ACL-Workshop on Multiword Expressions: Analysis, Acquisition, and Treatment, 65-72.

Bannard, C. (2005). 'Learning About the Meaning of Verb Particle Constructions from Corpora', Journal of Computer Speech and Language 19(4): 467-478.

Barkema, H. (1994). 'Determining the Syntactic Flexibility of Idioms', in U. Fries, G. Tottie and P. Schneider (eds.), Creating and Using English Language Corpora. Amsterdam: Rodopi, 39-52.

Barsalou, L. R. (1992). Cognitive Psychology: An Overview for Cognitive Scientists. Hillsdale, NJ: Erlbaum.

Berry-Rogghe, G. L. M. (1974). 'Automatic Identification of Phrasal Verbs', in Mitchell (ed.), Computers in the Humanities. Edinburgh: Edinburgh University Press, 16-26.

Birner, B. J. and Ward, G. (1998). Information Status and Noncanonical Word Order in English. Amsterdam/Philadelphia: John Benjamins.

Booij, G. (2010). Construction Morphology. Oxford: Oxford University Press.

Cacciari, C., and Glucksberg, S. (1991). 'Understanding Idiomatic Expressions: The Contribution of Word Meanings', in G. Simpson (ed.), Understanding Word and Sentence. The Hague: North Holland, 217-240.

Cacciari, C., and Tabossi, P. (eds.). (1995). Idioms: Processing, Structure, and Interpretation. Hillsdale, NJ: Lawrence Erlbaum.

Chafe, W. L. (1968). 'Idiomaticity as an Anomaly in the Chomskyan Paradigm', Foundations of Language 4(2): 109-127.

Chomsky, N. A. (1965). Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Collins Cobuild Dictionary of Idioms. 2002. 2nd ed. London: Harper Collins.

Cowie, A. P., and Howarth, P. (1996). 'Phraseology – A Select Bibliography', International Journal of Lexicography 9(1): 38–51.

Croft, W., and Cruse, D.A. (2004). Cognitive Linguistics. Cambridge: Cambridge University Press.

Dong, Q. P. (1971). 'The Applicability of Transformation to Idioms', CLS 7: 198-205.

Evans, V., and Green, M. (2006). Cognitive Linguistics: An Introduction. Mahwah, NJ/London: Lawrence Erlbaum.

Fernando, C. (1996). Idioms and Idiomaticity. Oxford: Oxford University Press.

Fillmore, C. J, Kay, P., and O'Connor, M. C. (1988). 'Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*', Language 64(3): 501-538.

Fraser, B. (1970). 'Idioms within a Transformational Grammar', Foundations of Language 6(1): 22-42.

Gibbs, R W., and Gonzales, G. P. (1985). 'Syntactic Frozenness in Processing and Remembering Idioms', Cognition 20(3): 243-259.

Gibbs, R. W., and Nayak, N. (1989). 'Psycholinguistic Studies on the Syntactic Behavior of Idioms', Cognitive Psychology 21(1): 100-138.

Gibbs, R. W., Nayak, N., Bolton, J., and Keppel, M. (1989). 'Speakers' Assumptions about the Lexical Flexibility of Idioms', Memory and Cognition 17(1): 58-68.

Glucksberg, S. (1993). 'Idiom Meanings and Allusional Content', in C. Cacciari and P. Tabossi (eds.), Idioms: Processing, Structure, and Interpretation. Hillsdale, NJ: Erlbaum, 3-26.

Goldberg, A. E. (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago: Chicago University Press.

Goldberg, A. E. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.

Gries, St. Th., and Stefanowitsch, A. (2004). Co-varying Collexemes in the *Into*-causative', in M. Achard and S. Kemmer (eds.), Language, Culture, and Mind. Stanford, CA: CSLI, 225-236.

Hymes, D. (1962). 'The Ethnography of Speaking', in T. Gladwin and W. C. Sturtevant (eds), Anthropology and Human Behavior. Washington, DC: The Anthropological Society of Washington.

Jackendoff, R. (1997). 'Twistin' the Night away', Language 73(3): 534-559.

Kay, P., and Fillmore, C J. (1999). 'Grammatical Constructions and Linguistic Generalizations: The *What's* X *Doing* Y? Construction', Language 75(1): 1-33.

Lakoff, G. (1987). Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: Chicago University Press.

Lambrecht, K. (1988). 'There Was a Farmer Had a Dog: Syntactic Amalgams Revisited', BLS 14: 319-339.

Lambrecht, K. (1990). '"What me, Worry?" – 'Mad Magazine' Sentences Revisited', BLS 16: 215-228.

Langacker, R. W. (1987). Foundations of Cognitive Grammar. Volume I. Stanford, CA: Stanford University Press.

Lin, D. (1999). 'Automatic Identification of Noncompositional Phrases', in Proceedings of the 37th Annual Meeting of the ACL, College Park, USA, 317-324.

Makkai, A. (1972). Idiom Structure in English. The Hague: Mouton de Gruyter.

McCarthy, D, Keller, B., and Carroll, J. (2003). 'Detecting a Continuum of Compositionality in Phrasal Verbs', in Proceedings of the ACL-Workshop on Multiword Expressions: Analysis, Acquisition, and Treatment, 73-80.

McGlone, M. S., Glucksberg, S., and Cacciari, C. (1994). 'Semantic Productivity and Idiom Comprehension', Discourse Processes 17(2): 176-190.

Michaelis, L. A., and Lambrecht, K. (1996). 'Toward a Construction-based Theory of Language Functions: The Case of Nominal Extraposition', Language 72(2): 215-247.

Moon, R. (1998). Fixed Expressions and Idioms in English: A Corpus-based Approach. Oxford: Clarendon.

Nattinger, J. R., and DeCarrico, J. S. (1992). Lexical Phrases and Language Teaching. Oxford: Oxford University Press.

Newmeyer, F. J. (1974). 'The Regularity of Idiom Behavior', Lingua 34: 327-42.

Nicolas, T. (1995). 'Semantics of Idiom Modification', in M. Everaert, E.-J. van der Linden, A. Schenk and R. Schreuder (eds.). Idioms: Structural and Psychological Perspectives. Hillsdale, NJ: Lawrence Erlbaum, 233-252.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). 'Idioms', Language 70(3): 491-538.

Pawley, A., and Syder, F. (1983). 'Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency', in J. C. Richards and R. W. Schmidt (eds.), Language and Communication. London: Longman, 191–226.

Peterson, R. R., and Burgess, C. (1993). 'Syntactic and Semantic Processing during Idiom Comprehension: Neurolinguistic and Psycholinguistic Dissociations', in C. Cacciari and

P. Tabossi (eds.), Idioms: Processing, Structure, and Interpretation. Hilldale, NJ: Erlbaum, 201-225.

Pierrehumbert, J. B. (2003). 'Probabilistic Phonology: Discrimination and Robustness', in R. Bod, J. Hay and S. Jannedy (eds.), Probabilistic Linguistics. Cambridge, MA: MIT Press, 177-228.

Prince, E. F. (1978). 'A Comparison of WH-clefts and *It*-clefts in Discourse. Language 54(4): 883-906.

Schone, P., and Jurafsky, D. (2001). 'Is Knowledge-free Induction of Multiword Unit Dictionary Headwords a Solved Problem?, in Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing, 100-108.

Speelman, D., Tummers, J., and Geeraerts, D. (2009). 'Lexical Patterning in a Construction Grammar: The Effect of Lexical Co-occurrence Patterns on the Inflectional Variation in Dutch Attributive Adjectives', Constructions and Frames 1(1): 87-118.

Stefanowitsch, A., and Gries, St. Th. (2003). 'Collostructions: Investigating the Interaction between Words and Constructions', International Journal of Corpus Linguistics 8(2): 209-243.

Titone, D. A., and Connine, C. M. (1994). 'The Comprehension of Idiomatic Expressions: Effects of Predictability and Literality', Journal of Experimental Psychology: Learning, Memory and Cognition 20(5): 1126-1138.

Wierzbicka, A. (1987). 'Boys will Be Boys', Language 63(1): 95-114.

Williams, E. (1994). 'Remarks on Lexical Knowledge', Lingua 92: 7-34.

Wray, A. (2002). Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Wulff, S. (2008). Rethinking Idiomaticity: A Usage-based Approach. London/New York: Continuum Press.

Zwicky, A. (1974). '"Hey whatsyourname!", CLS 10: 787-801.